Supplemental Experimental Procedures

Selection of RFAM families

To test whether ECs could predict tertiary structure contacts, we used RNA multiple sequence alignments from the RFAM 11.0 database (Burge et al., 2013) , removing columns with > 50% gaps. Mappings from the original RFAM coordinates to "RFAM reduced coordinates" – in which gaps have been removed – are recorded in Data. We restricted to families where the effective number of sequences (M_{eff}, see below) was greater than 0.5L, where L is the number of columns in the alignment, yielding 182 families (see https://marks.hms.harvard.edu/ev_rna/). Of these, 22 aligned to a known structure in the PDB (Berman et al., 2000). Ranked lists of ECs are contained in Data S1

Computing evolutionary couplings

Summary

To identify co-evolving nucleotides in RNA alignments, we fit a global statistical model to the sequences that is parameterized by single-site bias and pairwise coupling terms. In contrast to models of co-variation that consider pairs in isolation, such as mutual information (MI), this global statistical model can de-convolve transitive, chained co-variation into a typically smaller, more concentrated set of underlying couplings. In the following sections we (i) present the probability model; (ii) outline an approximate penalized Maximum Likelihood approach for fitting the model; and then describe three additional features that improve prediction accuracy, including (iii) regularization, (iv) sample reweighting; and (v) average product correction. Code is available at https://github.com/debbiemarkslab/plmc.

(i) Description maximum entropy probability model

We model the probability of a sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$ of length L as

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{L} h_i(\sigma_i) + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} J_{ij}(\sigma_i, \sigma_j)\right).$$

The external fields h_i represent single-site conservation and the pair couplings J_{ij} represent co-variation. For example, the term $J_{ij}(\sigma_i, \sigma_j)$ represents the statistical energy contributed by nucleotide σ_i in position i interacting with nucleotide σ_j in position j. Thus, if there are L nucleotides total, with each taking 5 possible states (A,C,G,U and gap), then $\boldsymbol{J}$ can be thought of as a $L \times L \times 5 \times 5$ matrix, where each 5×5 slice describes the pattern of co-variation between a given pair of positions.

The *partition function* Z ensures that P is properly normalized, and is given by

$$Z = \sum_{\sigma} \exp\left(\sum_{i=1}^{L} h_i(\sigma_i) + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} J_{ij}(\sigma_i, \sigma_j)\right)$$

Once the parameters $\boldsymbol{h}$ and $\boldsymbol{J}$ have been fit to data, we use the Frobenius norm $FN(i,j)$ of the J_{ij} couplings to assess the strength of coupling between position i and j, as follows.

$$FN(i,j) = \| J_{ij} \|_2 = \sqrt{\sum_k \sum_l {J'}_{ij}(k,l)^2}$$

where $J_{ij}{}'$ is a centered version of J_{ij} with row and column means set to 0. The *FN* scores are used to generate evolutionary couplings (ECs), as described in part (v).

(ii) Model fitting by pseudo-maximum likelihood (PLM)

A standard, consistent method for inferring the parameters of probability models is maximum likelihood, where the parameters are chosen to maximize the probability of the observed data under the model. Direct maximum likelihood is ill suited to the model described above, since it requires computing Z directly, which is intractable. Instead of maximizing the likelihood, we instead maximize a surrogate function, the pseudolikelihood (Besag, 1975) which approximates the full likelihood for each sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$ by a product of conditional likelihoods for each site i:

$$P(\sigma_1, \ldots, \sigma_L \,|\boldsymbol{h},\boldsymbol{J}) \approx \prod_{i=1}^{L} P(\sigma_i \,|\boldsymbol{\sigma} \setminus \sigma_i\,, \boldsymbol{h},\boldsymbol{J})$$

By computing a likelihood for each site i while conditioning on the remainder of the sequence $(\boldsymbol{\sigma} \setminus \sigma_i)$, the global partition function Z is replaced by a number of local partition functions, so that all terms in the approximate likelihood function (shown below) are tractable.

$$P(\sigma_i \,|\boldsymbol{\sigma} \setminus \sigma_i\,, \boldsymbol{h},\boldsymbol{J}) = \frac{\exp\left(h_i(\sigma_i) + \sum_{j\neq i} J_{ij}(\sigma_i\,, \sigma_j)\right)}{\sum_a \exp\left(h_i(a) + \sum_{j\neq i} J_{ij}(a, \sigma_j)\right)}$$

This pseudolikelihood approach has previously been applied (Ekeberg et al., 2013; Hopf et al., 2015; Hopf et al., 2014; Kamisetty et al., 2013; Ovchinnikov et al., 2014) to estimate residue couplings in protein sequence families. We optimize this approximate likelihood function (with some modifications outlined in (iii) and (iv)) using a quasi-Newton method (L-BFGS).

(iii) Regularization

The number of parameters to estimate in $\boldsymbol{J}$ is $\frac{L(L-1)}{2}q^2$, where L is the length of the sequence and q is the number of states (i.e. ~10^6 parameters for an RNA of length 200).

Since this vastly exceeds the effective number of sequences in a typical alignment, parameters must be strongly regularized to limit over-fitting. To that end, we use L_2-regularization of the fields $\boldsymbol{h}$ and couplings $\boldsymbol{J}$ with strength λ_h and λ_J respectively:

$$\mathcal{R}(\boldsymbol{h},\boldsymbol{J}) = \lambda_h \sum_{i=1}^{L} \| h_i \|_2^2 + \lambda_J \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \| J_{ij} \|_2^2$$

Optimizing this augmented objective involves a tradeoff between fitting the data (by increasing the pseudo-likelihood) and maintaining small parameter values (by decreasing the regularization term $\mathcal{R}$). If $\mathcal{L}(\boldsymbol{h},\boldsymbol{J})$ denotes the pseudo-likelihood of parameters $\boldsymbol{h}$ and $\boldsymbol{J}$ under the data, then we can write the optimization problem as follows.

$$\boldsymbol{h},\boldsymbol{J} = \arg\min_{\boldsymbol{h},\boldsymbol{J}} \left(-\log \mathcal{L}(\boldsymbol{h},\boldsymbol{J}) + \mathcal{R}(\boldsymbol{h},\boldsymbol{J})\right)$$

In this work, we set the regularization strength as $\lambda_h = 0.01$ and $\lambda_h = 20.0$ for all computations of ECs (i.e. both for RNA alone and RNA-protein together).

(iv) Sample reweighting

Our maximum entropy approach models the sequences in an alignment as independent draws from an underlying distribution. However, this assumption does not hold in reality, since sequences are usually related by phylogeny. To account for this, we reweight sequences in inverse proportion to the size of their sequence neighborhood. Formally, a sequence σ of length L in alignment A, is given the weight

$$w(\sigma) = \frac{1}{m(\sigma)} \quad \text{where} \quad m(\sigma) = |\{\sigma' \in A \mid \text{hamming}(\sigma,\sigma') * L^{-1} < \theta\}|$$

θ is a user-defined parameter determining neighborhood size. In this study, we used $\theta = 0.2$ for all RFAM alignments and $\theta = 0.1$ for RNA-protein alignments, except for alignments involving the HIV genes Rev and RRE, for which used $\theta = 0.033$.

The sum of weights $w(\sigma)$ over all sequences in the alignment represents the total *effective* number of sequences (M_{eff}). In other words,

$$M_{eff}(A) = \sum_{\sigma \in A} w(\sigma)$$

(v) Average product correction

The *FN* scores defined in part (i) cannot be directly used for inferring structure contacts, since they are contaminated by bias due to phylogeny and undersampling. Fortunately, these artifacts are concentrated in the top eigenvector of the *FN* matrix, and can therefore be removed using an average product correction (APC) , which reconstitutes the *FN* matrix from its eigenvectors while setting the top eigenvalue to 0. The resulting APC-

corrected *FN* matrix contains the evolutionary coupling (EC) scores referenced throughout this paper.

In practice, we apply the APC by subtracting normalized row and column averages from each position, as follows.

$$EC(i,j) = FN(i,j) - \frac{\left(\sum_{i' \neq i} FN(i',j)\right)\left(\sum_{j' \neq j} FN(i,j')\right)}{\sum_{i'} \sum_{j' \neq i} FN(i',j')}$$

Prediction of mutational effects

Our predictions of mutations likely to disrupt key interactions – presented for the T box riboswitch and RNase P – are based on the inferred parameters $\boldsymbol{h}$ and $\boldsymbol{J}$ of the global probability model (see above) and follow the methodology presented in (Hopf et al., 2015). Briefly, consider a sequence σ and mutation m. Let σ' be the sequence that results from applying m to σ. The effect of mutation m is calculated as

$$\text{Effect}(m) = E(\sigma') - E(\sigma) \quad \text{where} \quad E(\sigma) = \sum_{i=1}^{L} h_i(\sigma_i) + \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} J_{ij}(\sigma_i, \sigma_j)$$

Computing MI

To investigate how ECs compare to previous measures of co-evolution, we computed two versions of mutual information (MI). First we computed the raw MI (MI_R) as shown below, where $f_i(A) = P(S_i = A)$ and $f_{ij}(A,B) = P(S_i = A,\ S_j = B)$ for a sequence S in the alignment.

$$MI_R(i,j) = \sum_{A,B} \frac{f_{ij}(A,B)}{f_i(A) f_j(B)}$$

EC scores differ from MI_R in three ways: (1) They rely on a global maximum entropy model; (2) They down-weight sequences with a greater phylogenetic representation in the alignment; (3) They include an APC correction. Since feature (1) is the focus of this study, we also computed an enhanced MI score (MI_E), which incorporates features (2) and (3), as has been done in previous work on RNA co-evolution (Dunn et al., 2008).

Annotating interactions

For each alignment, we investigated the top L/2 contacts with chain-distance > 4. We first classified contacts as true-positives if the minimum-atom-distance from the crystal structure was < 8 Å. These were classified according secondary structure distance (d_{ss}) and biochemical interaction type, with long-range contacts defined as those satisfying d_{ss} > 4. The d_{ss} for a pair of bases is the length of the shortest path between them in a graph

where nodes are bases and edges are either secondary-structure contacts or instances of adjacency on the chain. To compute d_{ss}, we used the consensus secondary-structure provided by RFAM, which is inferred using a profile stochastic context-free grammar (Nawrocki and Eddy, 2013). To classify contacts by their biochemical interaction type, we used crystal structure annotations from FR3D (Petrov et al., 2011; Sarver et al., 2008) which were downloaded from RNA3DHub (http://rna.bgsu.edu/rna3dhub/). Ranked lists of ECs for each of the 22 RFAM families with a matching PDB structure are provided in Data S1.

Computing 3D structures from evolutionary couplings

We performed blinded structure prediction for all RNA families that (i) Have a known structure (ii) Have length between 70-120nt (iii) Have at least one *highly-long-range* contact, defined as a contact with $d_{ss} \geq L/4$, where L is the length of the RNA. We performed structure prediction with Nucleic Acid Simulation Tool (NAST) (Jonikas et al., 2009), a coarse-grained modeler that uses a combination secondary structure and tertiary contacts as inputs, followed by refinement in XPLOR (Schwieters et al., 2003). We describe the folding pipeline in detail below.

1) For each RNA family, we generated 200 random unfolded structures that satisfied the secondary structure constraints (Figure S7A).
2) Next, we performed molecular dynamics using tertiary structure restraints to generate candidate models (Figure S7B) with a restraint energy of 40. To obtain these tertiary structure restraints, we used the N long-range contacts with the top EC scores, where N varied between 20 and the length L of the RNA in intervals of 20. Thus, for a typical RNA family, we used around 4 different restraint sets, where the first set had the least contacts and the fourth had the most.
3) Since many restraint sets contained false-positives (i.e. restraints that are not satisfied in the true 3D structure), we used an iterative pruning procedure to remove contacts that were not consistent with the rest of the set. To that end, we performed molecular dynamics using weak constraints to iteratively remove restraints that were consistently violated by the resulting structures (Figure S7C), removing at most 15% of the contacts in any one round. Contacts were defined to be violated when the average distance between the corresponding bases was > 15 Å. At the end of this process, each restraint set from part (2) had been replaced by a subset, where all the restraints in the subset were consistent with each other.
4) At the end of steps (1-3), we obtained 200 decoy models for each restraint set, meaning 600-1000 decoys for each RNA family (longer RNAs had more restraint sets and therefore more decoys). We then assigned to each decoy an energy-per-contact, defined as E/N where E is NAST energy of the decoy and N is the number of contacts in its restraint set. For each RNA family, we then clustered the 20% of decoys with the lowest energy-per- contact using k-means with k = 4. RMSDs were calculated using Biopython, (Figure S7D). See Figure S4 for plots of energy vs. RMSD.

5) From each cluster, we chose a lowest energy representative and then created an all-atom structure using the NAST C2A pipeline (Figure S7E).
6) Finally, we refined the all-atom models by simulated annealing with XPLOR (Figure S7F). Thus, at the end of this pipeline, we produce four candidate predicted structures for each RNA family.

To analyze the predicted structures for each RNA family, we calculated the all-atom RMSD from the true structure using Pymol and used Molprobity (Davis et al., 2004) to analyze the produce scores quantifying the structures' intrinsic quality, i.e. how much they reproduce the geometry of 'typical' RNA structures. The predicted structure with lowest RMSD to the crystal structure is shown in Figure 3. In addition, we ran the above pipeline with no tertiary restraints as a control (see Figure 3B). RMSD values and Molprobity scores are available in Data S2.

RNA-protein 3D structure prediction

Selection of RNA-protein complexes

To compute evolutionary couplings between a pair of interacting genes, one must accurately phase the corresponding alignments (i.e. match sequences from one alignment with sequences from the other). Previous work detecting evolutionary couplings in protein-protein complexes (Hopf et al., 2014) has benefitted from co-operonic interaction partners, which can be accurately phased using genomic distance. Since RNA-protein complexes are typically not co-operonic, we limited our analysis to universally conserved, highly specific interactions between RNAs and proteins with no close paralogs. After excluding complexes with low sequence diversity (M_{eff} / L > 0.25) and those that share an interface with a third interaction partner, we arrived at a final validation set of 21 RNA-protein complexes (see Data S3 for ranked EC lists).

Obtaining protein alignments

RNA alignments for all RNA-protein complexes were taken from RFAM (Burge et al., 2013). Protein alignments were taken from PFAM (Finn et al., 2014) where it covers the full protein e.g. for RNaseP protein, and otherwise created by searching the UniProt (UniProt, 2015) database (release 2015_02) using 5 iterations of jackhammer (Finn et al., 2011), using an e-value, including columns with less than 50% gaps and removing sequences that had <50% length coverage relative to our query sequence. All resulting protein alignments are provided in Data S7.

Concatenation

Detecting coevolution in an RNA-protein complex requires phasing or "concatenating" the sequences in the alignments of the RNA and protein respectively. We used the NCBI taxonomy ID to concatenate RNA and protein sequences, as follows. First, we identified the set of NCBI taxonomy IDs with at least one representative in both the RNA and protein alignments. Next, for IDs with more than one RNA or protein representative, we computed the average hamming distance between representatives and discarded taxonomy IDs for which the average hamming distance exceeded 1%. Thus, the remaining taxonomy IDs each had one or more highly similar RNA representatives, and

one or more highly similar protein representatives. For each of these remaining IDs, we randomly chose an RNA representative and a protein representative, which together formed a line in the final alignment.

Calculating ECs

To compute RNA-protein ECs, we used the same approach as for RNA (described earlier in Methods) but now with a full alphabet including all amino acids. No other changes were made to the model.

Rigid body docking

To determine whether EC-derived RNA-protein contacts improve 3D structure prediction of RNA-protein complexes, we used these contacts as restraints for rigid body docking in HADDOCK (Dominguez et al., 2003). Specifically, we docked the 6 (out of 21) RNA-protein complexes that had least 75% true positives for the top 4 contacts. We inputted these top 4 contacts as unambiguous distance restraints (5$\pm$2 Å) into HADDOCK, and otherwise used default parameters. For docking controls, we applied center of mass restraints only. By default, HADDOCK clusters docked decoys and ranks the representatives of each cluster. We extracted the highest-ranking representative from each cluster for downstream analysis. All input models, restraint files, and output cluster representatives for cases and controls are provided in Data S7.

Computing i-RMSD

Interface-RMSD (i-RMSD) is a commonly used metric for assessing prediction accuracy of molecular complexes. I-RMSD is equivalent to standard RMSD, but taken over the interface between subunits, defined as the subset of atoms that lie within 10 Å (where distance is computed with respect to the experimental structure) of the subunit that they are not directly part of. We used Biopython to identify interface atoms and computed all-atom RMSDs in pymol (with no atom rejection).

Evolutionary couplings for HIV Rev Response Element (RRE)

We computed evolutionary couplings RRE using the RFAM alignment (RF00036) and also a custom alignment (referred to here as LANL) using sequences from Los Alamos HIV sequence database http://www.hiv.lanl.gov. To form the custom alignment, we downloaded *env* nucleotide sequences (aligned to the reference HXB2 genome). We then realigned these sequences with cmalign (Nawrocki and Eddy, 2013) using the cm profile from the reference RRE RFAM entry. *Rev* sequences (also downloaded from http://www.hiv.lanl.gov) were realigned by iterative alignment with hmmalign (Eddy, 1998) using the --hand option. To infer Rev-RRE inter-molecular contacts, we phased the LANL RRE and Rev alignments by matching sequences with same Genbank ID.

The RNA secondary structure of the SL4 RRE conformation originally proposed in (Mann et al., 1994) and the SL5 conformation were taken from (Sherpa et al., 2015) using the pNL4-3 (Genbank AF324493) genome. Energy calculations were performed using RNAeval (Lorenz et al., 2011; Mathews et al., 2004; Turner and Mathews, 2010) with the default settings from the online server http://rna.tbi.univie.ac.at/cgi-bin/RNAeval.cgi.

Evolutionary couplings for T box riboswitch and RNase P

Contacts for the T box riboswitch (RF00230) and RNase P family members (RF00010, RF00009 and RF00373) – presented in Figure 7 – were drawn from the top L ECs with a chain distance > 4. Given the large size of these sequences, we defined contacts as long-range when they satisfied $d_{ss} \geq 12$. For the archeal RNase P (RF00373), these criteria produced a set of relatively low ranking contacts that appeared in isolation at apparently random positions in the contact map, a hallmark of false-positives. Therefore, for RF00373, we removed contacts with rank > L/2 unless they were reinforced by at least one other contact, where two contacts are considered mutually reinforcing if their endpoints are both within 1 bp of each other. Contacts mentioned in the text are given in terms of their RFAM-reduced coordinates unless stated otherwise. The six long-range contacts referred to in the discussion of the T box riboswitch are (91,191), (92,192), (91,186), (90,210), (91,210) and (90,186) in RFAM reduced numbering.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.
Besag, J. (1975). Statistical analysis of non-lattice data. The statistician, 179-195.
Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., and Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. Nucleic Acids Res *41*, D226-232.
Davis, I.W., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2004). MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res *32*, W615-619.

Dominguez, C., Boelens, R., and Bonvin, A.M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc *125*, 1731-1737.
Dunn, S.D., Wahl, L.M., and Gloor, G.B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics *24*, 333-340.
Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.
Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys *87*, 012707.
Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J.*, et al.* (2014). Pfam: the protein families database. Nucleic Acids Res *42*, D222-230.
Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res *39*, W29-37.
Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Springer, M., Sander, C., and Marks, D.S. (2015). Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv preprint arXiv:151004612.
Hopf, T.A., Scharfe, C.P., Rodrigues, J.P., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife *3*.
Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D., and Altman, R.B. (2009). Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA *15*, 189-199.
Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A *110*, 15674-15679.
Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26.
Mann, D.A., Mikaelian, I., Zemmel, R.W., Green, S.M., Lowe, A.D., Kimura, T., Singh, M., Butler, P.J., Gait, M.J., and Karn, J. (1994). A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. J Mol Biol *241*, 193-207.
Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., and Turner, D.H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A *101*, 7287-7292.
Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933-2935.
Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife *3*, e02030.
Petrov, A.I., Zirbel, C.L., and Leontis, N.B. (2011). WebFR3D--a server for finding, aligning and analyzing recurrent RNA 3D motifs. Nucleic Acids Res *39*, W50-55.

Sarver, M., Zirbel, C.L., Stombaugh, J., Mokdad, A., and Leontis, N.B. (2008). FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J Math Biol *56*, 215-252.
Schwieters, C.D., Kuszewski, J.J., Tjandra, N., and Clore, G.M. (2003). The Xplor-NIH NMR molecular structure determination package. J Magn Reson *160*, 65-73.
Sherpa, C., Rausch, J.W., Le Grice, S.F., Hammarskjold, M.L., and Rekosh, D. (2015). The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication. Nucleic Acids Res *43*, 4676-4686.
Turner, D.H., and Mathews, D.H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res *38*, D280-282.
UniProt, C. (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.